A deep learning framework for neuroscience

Blake A. Richards*[1,2,3,4], Timothy P. Lillicrap*[5,6,], Philippe Beaudoin[7], Yoshua Bengio[1,4,8], Rafal Bogacz[9], Amelia Christensen[10], Claudia Clopath[11], Rui Ponte Costa[12,13], Archy de Berker[7], Surya Ganguli[14,15], Colleen J. Gillon[16], Danijar Hafner[15,18,19], Adam Kepecs[20], Nikolaus Kriegeskorte[21,22], Peter Latham[22], Grace W. Lindsay[22,24], Ken Miller[22,24,25], Richard Naud[26,27], Christopher C. Pack[3], Panayiota Poirazi[28], Pieter Roelfsema[29], João Sacramento[30], Andrew Saxe[31], Benjamin Scellier[1,8], Anna Schapiro[32], Walter Senn[13], Greg Wayne[5], Daniel Yamins[33,34,35], Friedemann Zenke[36,37], Joel Zylberberg[4,38,39], Denis Therien*[7], Konrad P. Kording*[4,40,41]

1: Mila, Montréal, QC, Canada
2: School of Computer Science, McGill University, Montréal, QC, Canada
3: Department of Neurology & Neurosurgery, McGill University, Montréal, QC, Canada
4: Canadian Institute for Advanced Research, Toronto, ON, Canada
5: DeepMind, Inc., London, UK
6: Centre for Computation, Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London, UK
7: Element AI, Montréal, QC, Canada
8: Université de Montréal, Montréal, QC, Canada
9: MRC Brain Network Dynamics Unit, University of Oxford, Oxford, UK
10: Department of Electrical Engineering, Stanford University, Stanford, CA, USA
11: Department of Bioengineering, Imperial College London. UK
12: Computational Neuroscience Unit, School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths, University of Bristol, Bristol, UK
13: Department of Physiology, Universität Bern, Bern, Switzerland
14: Department of Applied Physics, Stanford University, Stanford, CA, USA
15: Google Brain, Mountain View, CA, USA
16: Department of Biological Sciences, University of Toronto Scarborough, Toronto, ON, Canada,
17: Department of Cell & Systems Biology, University of Toronto, Toronto, ON, Canada
18: Department of Computer Science, University of Toronto, Toronto, ON, Canada
19: Vector Institute, Toronto, ON, Canada
20: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
21: Department of Psychology and Neuroscience, Columbia University, New York, NY, USA
22: Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA
23: Gatsby Computational Neuroscience Unit, University College London, London, UK
24: Center for Theoretical Neuroscience, Columbia University, New York, NY, USA
25: Department of Neuroscience, College of Physicians and Surgeons, Columbia University. New York, NY, USA
26: University of Ottawa Brain and Mind Institute, Ottawa, ON, Canada,
27: Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, ON, Canada
28: Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), Crete, Greece
29: Department of Vision & Cognition, Netherlands Institute for Neuroscience, Amsterdam, The

Netherlands
30: Institute of Neuroinformatics, ETH Zürich and University of Zürich, Zürich, Switzerland
31: Department of Experimental Psychology, University of Oxford, Oxford, UK
32: Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA
33: Department of Psychology, Stanford University, Stanford, CA, USA
34: Department of Computer Science, Stanford University, Stanford, CA, USA
35: Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA
36: Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland
37: Centre for Neural Circuits and Behaviour, University of Oxford, Oxford, UK
38: Department of Physics and Astronomy York University, Toronto, ON, Canada
39: Center for Vision Research, York University, Toronto, ON, Canada
40: Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
41: Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA

* These authors contributed equally to this work.

Abstract
Systems neuroscience seeks explanations for how the brain implements a wide variety of perceptual, cognitive and motor tasks. Conversely, artificial intelligence attempts to design computational systems based on the tasks they will have to solve. In the case of artificial neural networks, the three components specified by design are the objective functions, the learning rules, and architectures. With the growing success of deep learning, which utilizes brain-inspired architectures, these three designed components have increasingly become central to how we model, engineer and optimize complex artificial learning systems. Here we argue that a greater focus on these components would also benefit systems neuroscience. We give examples of how this optimization-based framework can drive theoretical and experimental progress in neuroscience. We contend that this principled perspective on systems neuroscience will help to generate more rapid progress.

Introduction

Major technical advances are revolutionizing our ability to observe and manipulate brains at a large-scale and quantify complex behaviors[1,2]. How should we use this data to develop models of the brain? When the classical framework for systems neuroscience was developed, we could only record from small sets of neurons. In this framework, a researcher observes neural activity, develops a theory of what individual neurons compute, then assembles a circuit-level theory of how the neurons combine their operations. This approach has worked well for simple computations. For example, we know how central pattern generators control rhythmic movements[3], how the vestibulo-ocular reflex promotes gaze stabilization[4], and how the retina computes motion[5]. But, can this classical framework scale up to recordings of thousands of neurons and all of the behaviors that we may wish to account for? Arguably, we have not had as much success with the classical approach in large neural circuits that perform a multitude of functions, like the neocortex or hippocampus. In such circuits, researchers often find neurons with response properties that are difficult to summarize in a succinct manner[6,7].

The limitations of the classical framework suggest that new approaches are needed to take advantage of experimental advances. A promising framework is emerging from the interactions between neuroscience and Artificial Intelligence (AI)[8–10]. The rise of deep learning as a leading machine learning method invites us to revisit Artificial Neural Networks (ANNs). At their core, ANNs model neural computation using simplified units that loosely mimic the integration and activation properties of real neurons[11]. Units are implemented with varying degrees of abstraction, ranging from highly simplified linear operations to relatively complex models with multiple compartments, spikes, etc.[11–14]. Importantly, the *specific computations* performed by ANNs are not designed, but learned[15].

However, human design still plays a role in determining three essential components in ANNs: the learning goal, expressed as an objective function (or loss function) to be maximized or minimized; a set of learning rules, expressed as synaptic weight updates; and the network architecture, expressed as the pathways and connections for information flow (**Fig. 1**)[15]. Within this framework, we do not seek to summarize how a computation is performed, but we do summarize what objective functions, learning rules and architectures would enable learning of that computation.

Deep learning can be seen as a rebranding of long-standing ANN ideas[11]. Deep ANNs possess multiple layers, either feedforward, or recurrent over time. The “layers” are best thought of as being analogous to brain regions, rather than as specific laminae in biological brains[16,17]. “Deep” learning specifically refers to training hierarchical ANNs in an end-to-end manner, such that plasticity in each layer of the hierarchy contributes to the learning goals[15], which requires a solution to the “credit assignment problem” (Box 1)[18,19]. In recent years, progress in deep learning has come from the use of bigger ANNs, trained with bigger datasets using Graphics Processing Units (GPUs) that can efficiently handle the required computations. Such developments have produced solutions for many new problems, including image[20] and speech[21] classification and generation, language processing and translation[22], haptics and grasping[23], navigation[24], sensory prediction[25], game playing[26] and reasoning[27].

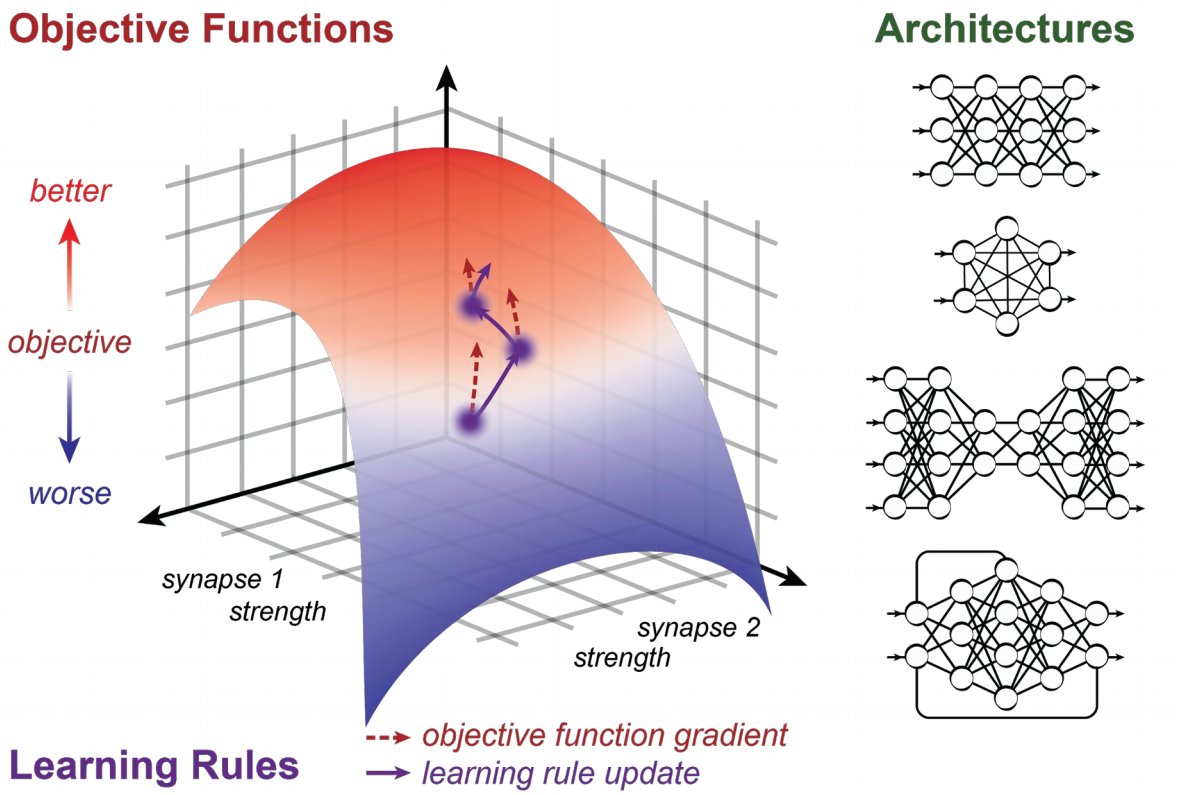

Figure 1: The three core components of ANN design. When designing ANNs, researchers do not craft the specific computations performed by the network. Instead they specify these three components. Objective functions quantify the performance of the network on a task, and learning involves finding synaptic weights that maximize or minimize the objective function. (Often, these are referred to as "loss" or "cost" functions.) Learning rules provide a recipe for updating the synaptic weights. This can lead to ascent of the objective, even if the explicit gradient of the objective function isn't followed. Architectures specify the arrangement of units in the network, and determine the flow of information, as well as the computations that are or are not possible for the network to learn.

Many recent findings suggest that deep learning can inform our theories of the brain. First, it has been shown that deep ANNs can, in some cases closely, mimic the representational transformations in primate perceptual systems[17,28], and thereby can be leveraged to manipulate neural activity[29]. Second, many well-known behavioral and neurophysiological phenomena, including grid cells[24], shape tuning[30], temporal receptive fields[31], visual illusions[32], and apparent model-based reasoning[33], have been shown to emerge in deep ANNs trained on tasks similar to those solved by animals. Third, many modeling studies have demonstrated that the apparent biological implausibility of end-to-end learning rules, e.g. learning algorithms that can mimic the power of the canonical backpropagation-of-error algorithm (backprop) (see Box 1), is overstated. Relatively simple assumptions about cellular and subcellular electrophysiology, inhibitory microcircuits, patterns of spike timing, short term plasticity, and feedback connections can enable biological systems to approximate backprop-like learning in deep ANNs[12,14,34–39]. Hence, ANN-based models of the brain may not be as unrealistic as previously thought, and simultaneously, they appear to explain a lot of neurobiological data.

With these developments, it is the right time to consider a deep-learning-inspired framework for systems neuroscience[8,19,40]. We have a growing understanding of the key principles that underlie ANNs, and there are theoretical reasons to believe that these insights

Box 1: Learning and the "credit assignment problem"

A natural definition of learning is that it is a change to a system that improves its performance. Suppose we have an objective function, F(W), which measures how well a system is currently performing, given the N-dimensional vector of its current synaptic weights, W. If the synaptic weights change from W to W + ΔW, then the change in performance is ΔF = F(W + ΔW) −F(W). If we make small changes to W, and F is locally smooth, then ΔF is given approximately by

$$\Delta F \approx \Delta W^{T} \cdot \nabla_{W} F$$

where $\nabla_W F$ is the gradient of F with respect to W[41]. Suppose we want to guarantee improved performance, i.e. we want to ensure ΔF > 0. We know that there is an N-1 dimensional manifold of local changes in W that all lead to the same improvement. Which one should we choose? Gradient-based algorithms derive from the intuition that we want to take the smallest step that gets us a specific level of improvement. If we choose a small step size, η, times the gradient $\nabla_W F$, then we will improve as much as possible for that step size. Thus, we have:

$$\Delta F \approx \eta \nabla_{W} F^{T} \cdot \nabla_{W} F > 0$$

In other words, the objective function value increases with every step (when η is small) according to the length of the gradient vector.

The concept of "credit assignment" refers to the problem of determining how much "credit" or "blame" a given neuron or synapse should get for a given outcome. More specifically, it is a way of determining how each parameter in the system (e.g., each synaptic weight) should change to ensure that ΔF > 0. In its simplest form, the "credit assignment problem" refers to the difficulty of assigning credit in complex networks. Updating weights using the gradient of the objective function, $\nabla_W F$, has proven to be an excellent means of solving the credit assignment problem in ANNs. A question that systems neuroscience faces is whether the brain also approximates something like gradient-based methods.

The most common method for calculating gradients in deep ANNs is backprop[15]. It uses the chain rule to recursively calculate gradients backwards from the output[11]. But backprop rests on biologically implausible assumptions, such as symmetric feedback weights and distinct forward and backward passes of information[14]. Many different learning algorithms, not just backprop, can provide estimates of a gradient, and some of these do not suffer from backprop's biological implausibility[12,14,34–38,91-93]. However, algorithms differ in their variance and bias properties (**Fig. 2**)[36,94]. Algorithms such as weight/node perturbation, which reinforce random changes in synaptic weights through rewards, have high variance in their path along the gradient[94]. Algorithms that use random feedback weights to communicate gradient information have high bias[36,95]. Various proposals have been made to minimize bias and variance in algorithms while maintaining their biological realism[37,38].

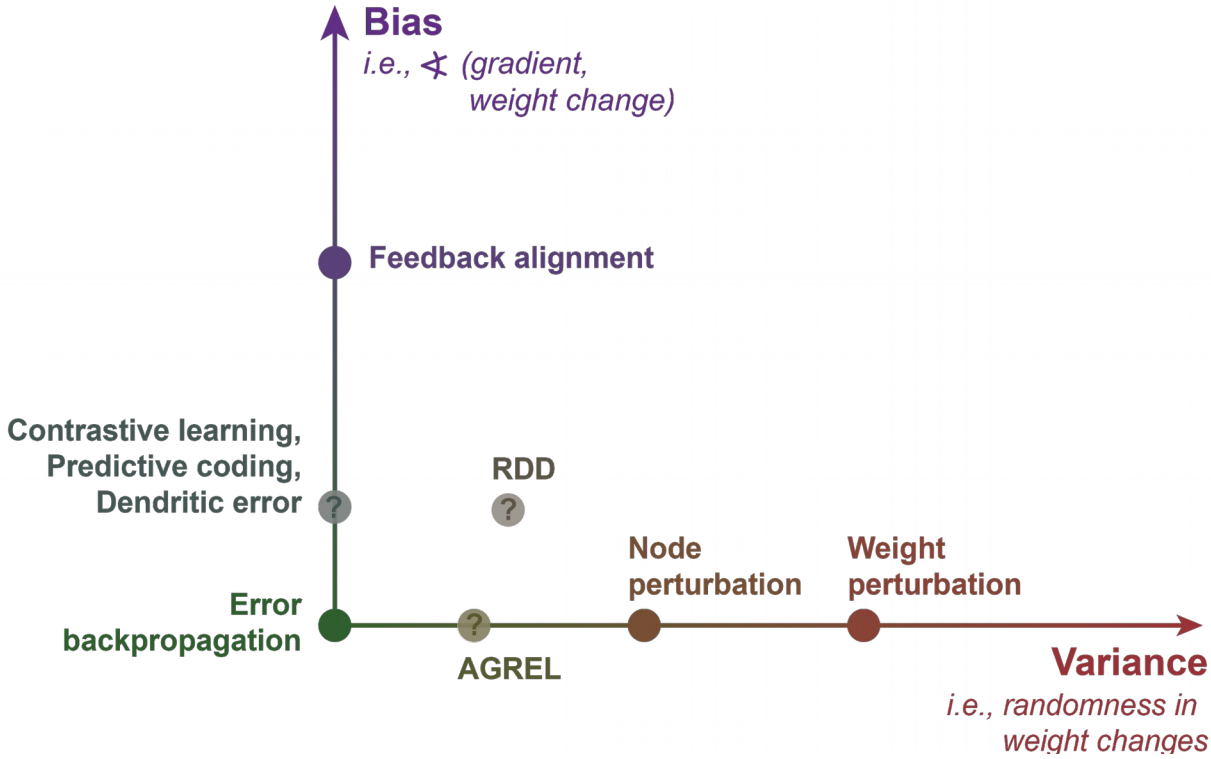

Fig 2: Bias and variance in learning rules. Many learning rules provide an estimate of the gradient of an objective function, even if they are not explicitly gradient-based. However, as with any estimator, these learning rules can exhibit different degrees of variance and bias in their estimates of the gradient. Here, we provide a rough illustration of how much bias and variance some of the proposed biologically plausible learning rules may have relative to backprop. It is important to note that the exact bias and variance properties of many of the learning rules are unknown, and this is just a sketch. As such, for some of the learning rules shown here, e.g. contrastive Hebbian learning, predictive coding (ref. 35), dendritic error learning (ref. 14), regression discontinuity design (RDD) (ref. 93), and attention-gated reinforcement learning (AGREL) (ref. 37), we have indicated their location with a question mark. For others, namely backpropagation, feedback alignment (ref. 36), and node/weight perturbation (ref. 94), we show their known relative positions.

apply generally[41,42]. Concomitantly, our ability to monitor and manipulate large neural populations opens the door to new ways of testing hypotheses derived from the deep learning literature. Here we sketch the scaffolding of a deep learning framework for modern systems neuroscience.

Constraining learning in artificial neural networks and the brain with "task sets"

The "No Free Lunch Theorems" demonstrated broadly that no learning algorithm can perform well on all possible problems[43]. ANN researchers in the first decade of the 21st century thus argued that AI should be primarily concerned with the set of tasks that "...most animals can perform effortlessly, such as perception and control, as well as ... long-term prediction, reasoning, planning, and [communication]"[44]. This set of tasks has been termed the "AI Set", and the focus on building computers with capabilities that are similar to those of humans and animals is what distinguishes AI tasks from other tasks in computer science[44] (note that the word "tasks" here refers broadly to any computation, including those that are unsupervised.)

Much of the success of deep learning can be attributed to the consideration given to learning in the AI Set[15,44]. Designing ANNs that are well-suited to learn specific tasks is an

example of incorporating “inductive biases” (Box 2): assumptions that one makes about the nature of the solutions to a given optimization problem. Deep learning works so well, in part, because it uses appropriate inductive biases for the AI Set[15,45], particularly hierarchical architectures. For example, images can be well described by composing them into a hierarchical set of increasingly complex features: from edges, to simple combinations of these, to larger configurations that form objects. Language too can be considered a hierarchical construction, with phonemes assembled into words, words into sentences, sentences into narratives. However, deep learning also eschews hand-engineering, allowing the function computed by the system to emerge during learning[15]. Thus, despite the common belief that deep learning relies solely on increases in computational power, or that it represents a “blank slate” approach to intelligence, many of the successes of deep learning have grown out of a balance between useful inductive biases and emergent computation, echoing the blend of nature and nurture which underpins the adult brain.

Similarly, neuroscientists focus on the behaviors/tasks that a species evolved to perform. This set of tasks overlaps with the AI Set, though possibly not completely, since different species have evolved strong inductive biases for their ecological niches. By considering this “Brain Set” for specific species—the tasks that are important for survival and reproduction for that species—researchers can focus on the features most likely to be key to learning. Just as departing from a pure blank slate was the key to the success of modern ANNs—e.g. by focusing on ANN designs with inductive biases that are useful for the AI Set—so we suspect that it will also be crucial to the development of a deep learning framework for systems neuroscience to focus on how a given animal might solve tasks in its appropriate Brain Set.

Box 2: What are inductive biases?
Learning is easier when we have prior knowledge about the kind of problems that we will have to solve[43]. Inductive biases are a means of embedding such prior knowledge into an optimization system. Such inductive biases may be generic, such as hierarchy, or specific, such as convolutions. Importantly, the inductive biases that exist in the brain will have been shaped by evolution to increase an animal’s fitness in both the broad context of life on Earth (e.g. life in a three-dimensional world where one needs to obtain food, water, shelter, etc.), and in specific ecological niches. Examples of inductive biases are:

Simple explanations: When attempting to make sense of the world, simple explanations may be preferred, as articulated by Occam’s Razor[96]. We can build this into ANNs using either Bayesian frameworks or by other mechanisms, such as sparse representations[59].

Object permanence: The world is organized into objects, which are spatiotemporally constant. We can build this into ANNs by learning representations that assume consistent movement in sensory space[97].

Visual translation invariance: A visual feature tends to have the same meaning regardless of its location. We can build this into ANNs using convolution operations[98].

Focused attention: Some aspects of the information coming into a system are more important than others. We can build this into ANNs through attention mechanisms[99].

Recognizing the importance of inductive biases in deep learning also helps address some existing misconceptions. Deep networks are often considered different from brains because they depend on large amounts of data. However, it is worth noting that (1) many species, especially humans, develop slowly with large quantities of experiential data and (2) that deep networks can work well in low data regimes if they have good inductive biases[46]. For example, deep networks can learn how to learn quickly[47]. In the case of brains, evolution could be one means by which such inductive biases are acquired[48,49].

The three core components of a deep learning framework for the brain

Deep learning combines human design with automatic learning to solve a task. What is designed are not the computations (i.e. the specific input/output functions of the ANNs), but three components: (1) objective functions, (2) learning rules, and (3) architectures (**Fig. 1**). **Objective functions** describe the goals of the learning system. They are functions of the synaptic weights of a neural network and the data it receives, but they can be defined without making reference to a *specific* task or dataset. For example, the cross-entropy objective function, which is common in machine learning, specificies a means of calculating performance on *any* categorization task, from distinguishing different breeds of dog in the ImageNet dataset to classifying the sentiment behind a tweet. We will return to some of the specific objective functions proposed for the brain below[50–53]. **Learning rules** describe how the parameters in a model are updated. In ANNs, these rules are generally used to improve on the objective function. Notably, this is true not only for supervised learning (where an agent receives an explicit target to mimic), but also for unsupervised learning (where an agent must learn without any instruction) and reinforcement learning systems (where an agent must learn using only rewards/punishments). Finally, **architectures** describe how the units in an ANN are arranged and what operations they can perform. For example, convolutional networks impose a connectivity pattern whereby the same receptive fields are applied repeatedly over the spatial extent of an input.

Why do so many AI researchers now focus on objective functions, learning rules and architectures instead of designing specific computations? The short answer is that this appears to be the most tractable way to solve real-world problems. Originally, AI practitioners believed that intelligent systems could be hand-designed by piecing together elementary computations[54]. But results on the AI Set were underwhelming[11]. It now seems clear that solving complex problems with pre-designed computations (e.g. such as handcrafted features) is usually too difficult and practically unworkable. In contrast, specifying objective functions, architectures, and learning rules works well.

There is, though, a drawback: the computations that emerge in large-scale ANNs trained on high-dimensional datasets can be difficult to interpret. We can construct a neural network in a few lines of code, and for each unit in an ANN we can specify the equations that determine their responses to stimuli or relationships to behavior. However, after training, a network is characterized by millions of weights that collectively encode what the network has learned, and it is hard to imagine how we could describe such a system with only a small number of parameters, let alone in words[55].

Such considerations of complexity are informative for neuroscience. For small circuits comprising only tens of neurons it may be possible to build compact models of individual neural

responses and computations (i.e. to develop models that can be communicated using a small number of free parameters or words)[3–5]. But, considering that animals are solving many AI Set problems, it is likely that the brain uses solutions that are as complex as the solutions used by ANNs. This suggests that a normative framework that explains why neural responses are as they are, might be best obtained by viewing neural responses as an emergent consequence of the interplay between objective functions, learning rules, and architecture. With such a framework in hand, one could then train ANN models that do, in fact, predict neural responses well[29,67,68]. Of course, those ANN models would likely be non-compact, involving millions, billions or even trillions of free parameters, and being nigh indescribable with words. Hence, our claim is not that we could ever hope to predict neural responses with a compact model, but rather, that we could explain the emergence of neural responses within a compact framework.

A question that naturally arises is whether the environment, or data, that an animal encounters should be a fourth essential component for neuroscience. Determining the "Brain Set" for an animal necessarily involves consideration of its evolutionary and ontogenic milieu. Efforts to efficiently describe naturalistic stimuli and identify ethologically-relevant behaviors are crucial to neuroscience, and have shaped many aspects of nervous systems. However, the core issue we are addressing in this perspective piece is how to develop models of complex, hierarchical brain circuits, so we view the environment as a crucial consideration to anchor the core components, but not as one of the components itself.

Once the appropriate Brain Set has been identified, the first question is: what is the architecture of the circuits? This involves descriptions of the cell types and their connectivity (micro, meso and macroscopic). Thus, uncontroversially, we propose that circuit-level descriptions of the brain are a crucial topic for systems neuroscientists. Thanks to modern techniques for circuit tracing and genetic lineage determination, rapid progress is being made[56,57]. But, to reiterate, we would argue that understanding the architecture is not sufficient for understanding the circuit; rather, it should be complemented by knowledge of learning rules and objective functions.

Many neuroscientists recognize the importance of learning rules and architecture. But identifying the objective functions that have shaped the brain, either during learning or evolution, is less common. Unlike architectures and learning rules, objective functions may not be directly observable in the brain. Nonetheless, we can define them mathematically and without making reference to a specific environment or task. For example, predictive coding models minimize an objective function known as the description length, which measures how much information is required to encode sensory data using the neural representations. Several other objective functions have been proposed for the brain (Box 3). In this perspective piece, we are not advocating for any of these specific objective functions in the brain, as we are articulating a framework, not a model. One of our key claims is that even though we must infer them, objective functions are an attainable part of a complete theory of how the architectures or learning rules help to achieve a computational goal.

This optimization framework has an added benefit: as with ANNs, the architectures, learning rules and objective functions of the brain are likely relatively simple and compact, at least in comparison to the list of computations performed by individual neurons[58]. The reason is that these three components must presumably be conveyed to offspring through a limited information bottleneck, i.e. the genome (which may not have sufficient capacity to fully specify

the wiring of large vertebrate brains[48]). In contrast, the environment in which we live can convey vast amounts of complex and changing information that dwarf the capacity of the genome.

Box 3: Are there objective functions for brains?
Animals clearly have some baseline objective functions. For example, homeostasis minimizes an objective function corresponding to the difference between a physiological variable (like blood oxygen levels) and a set-point for that variable. Given the centrality of homeostasis to physiology, objective functions are arguably something that the brain must be concerned with.

But, some readers may doubt whether the sort of objective functions used in machine learning are relevant to the brain. For example, the cross-entropy objective function used in ANNs trained on categorization tasks is unlikely to be used in the brain, since it requires specification of the correct category for each sensory input. Other objective functions are more ecologically plausible, though. Examples include the description length objective function used in predictive coding models[50], the log-probability of action sequences scaled by the reward they have produced (which is used in reinforcement learning to maximize rewards)[51], increases in mutual information with the environment[100], and empowerment[52,53], which measures the degree of control an agent has in their environment. These objective functions can all be specified mathematically for the brain without worrying about specific datasets, tasks or environments.

There are, however, real challenges in tying objective functions to empirical and theoretical models in neuroscience. Many potential plasticity rules may not follow the gradient of any objective function at all, or only follow it partially (**Fig. 3**). This apparently complicates our problems, and makes it impossible to guarantee that objective functions are always involved in neural plasticity. As well, the brain likely optimizes multiple objective functions[40], some of which we may in fact learn (i.e. we may "learn-to-learn"; for example, humans learn how to learn new board games), and some of which may have been optimized over the course of evolution rather than in an individual animal (i.e. reflexes or reproductive behavior).

Despite these complexities, we believe that consideration of objective functions is critical for systems neuroscience. After all, we know that biological variables, such as dopamine release, meaningfully relate to objective functions from reinforcement learning[64]. In addition, although many potential learning rules may not directly follow the gradient of the objective function, they would still lead to an improvement in that objective function. Here, identifying an objective function allows us to establish whether a change in the phenotype of a neural circuit should be considered a form of learning. If things don't "get better" according to some metric, how can we refer to any phenotypic plasticity as "learning"as opposed to just "changes"?

Since the responses of individual neurons are shaped by the environment, their computations should reflect this massive information source. We can see evidence of this in the ubiquity of neurons in the brain that have high entropy in their activity and that do not exhibit easy-to-describe correlations with the multitude of stimuli and behaviors that experimentalists have explored to date[6,7]. To clarify our claim, we are suggesting that identifying a normative explanation using the three components may be a fruitful way to go on to develop better, non-compact models of the response properties of neurons in a circuit, as shown by recent studies that use task-optimized deep ANNs to determine the optimal stimuli for activating specific neurons[29]. As an analogy, the theory of evolution by natural selection provides a compact

explanation for why species emerge as they do, one which can be stated in relatively few words. This compact explanation of the emergence of species can then be used to develop more complex, non-compact models of the phylogeny of specific species. Our suggestion is that normative explanations based on the three components could provide similar high-level theories for generating our lower-level models of neural responses, and that this would bring us one step closer to the form of "understanding" that many scientists seek.

It is worth recognizing that researchers have long postulated objective functions and plasticity rules to explain the function of neural circuits[59–62]. Many of them, however, have sidestepped the question of hierarchical credit assignment, which is key to deep learning[15]. There are clear experimental success stories too, including work on predictive coding[31,63], reinforcement learning[64,65], and hierarchical sensory processing[17,28]. Thus, the optimization-based framework that we articulate here can, and has, operated alongside studies of individual neuron response properties. But, we believe that we will see even greater success if a framework focused on the three core components is adopted more widely.

Architectures, learning rules, and objective functions in the wet lab

How can the framework articulated here engage with experimental work? One way to make progress is to build working models using the three core components, then compare the models with the brain. Such models should ideally check out on all levels: (1) They should solve the complex tasks from the Brain Set under consideration. (2) They should be informed by our knowledge of anatomy and plasticity. And, (3) they should reproduce the representations, and changes in representation, we observe in brains (**Fig. 4**). Of course, checking each of these criteria will be non-trivial. It may require many new experimental paradigms. Checking that a model can solve a given task is relatively straightforward, but representational and anatomical matches are not straightforward to establish, and this is an area of active research[66,67]. Luckily, the modularity of the optimization framework allows researchers to attempt to study each of the three components in isolation.

Empirical studies of architecture in the brain

To be able to identify the architecture that defines the inductive biases of the brain, we need to continue performing experiments that explore neuroanatomy at the circuit level. To really frame neuroanatomy within an optimization framework, we must also be able to identify what information is available to a circuit, including where signals about action outcomes may come from. Ultimately, we want to be able to relate these aspects of anatomy to concrete biological markers that guide the developmental processes responsible for learning.

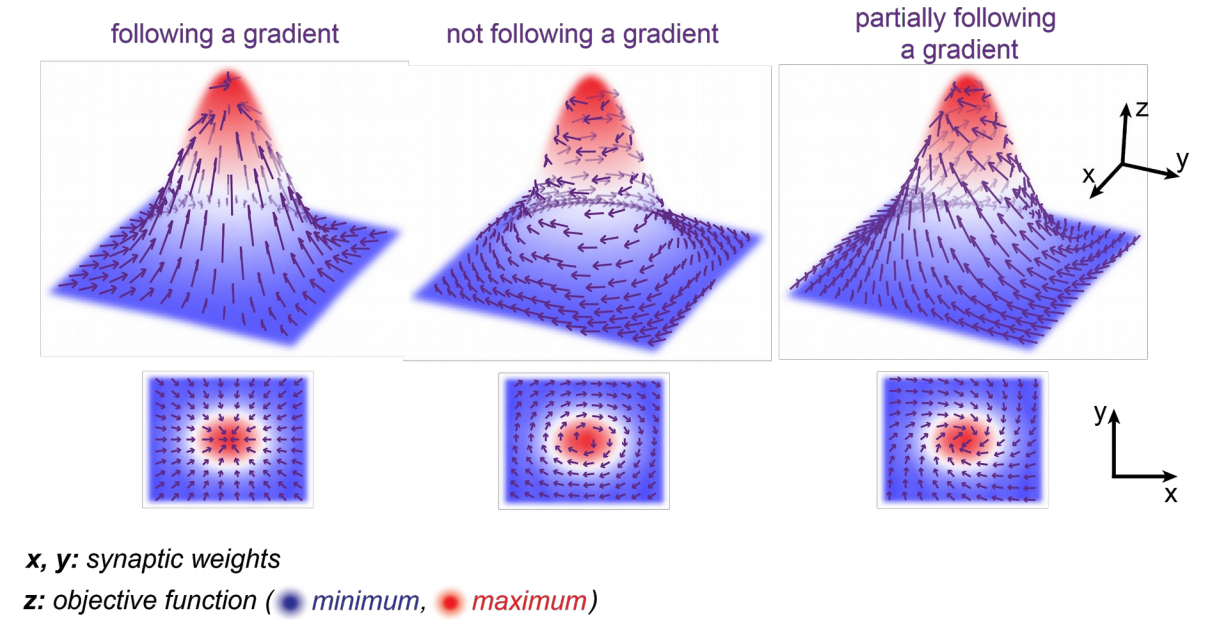

Figure 3: Learning rules that don't follow gradients. Learning should ultimately lead to some form of improvement, which could be measured with an objective function. But, not all synaptic plasticity rules need to follow a gradient. Here we illustrate this idea by showing three different hypothetical learning rules, characterized as vector fields in synaptic weight space. The x and y dimensions correspond to synaptic weights, and the z dimension corresponds to an objective function. Any vector field can be decomposed into a gradient and the directions orthogonal to it. On the left is a plasticity rule that adheres to the gradient of an objective function, directly bringing the system up to the maximum. In the middle is a plasticity rule that is orthogonal to the gradient, and as such, never brings the system closer to the maximum. On the right is a learning rule that only partially follows the gradient, bringing the system towards the maximum, but indirectly. Theoretically, any of these situations may hold in the brain, though learning goals would only be met in the cases where the gradient is fully or partially followed (left and right).

There is considerable experimental effort already underway towards describing the anatomy of the nervous system. We are using a range of imaging techniques to quantify the anatomy and development of circuits[57,68]. Extensive work is also conducted in mapping out the projections of neural circuits with cell-type specificity[56]. Research attempting to map out the hierarchy of the brain has long existed[69], but several groups are now probing which parts of deep ANN hierarchies may best reflect which brain areas[17,70]. For example, the representations in striate cortex (as measured, for example, by dissimilarity matrices) better match early layers of a deep ANN, while those in inferotemporal cortex better match later layers[8,71]. This strain of work also involves optimization of the architecture of deep ANNs so that they provide a closer fit to representation dynamics in the brain, e.g. by exploring different recurrent connectivity motifs[66]. Confronted with a bewildering set of anatomical observations that have been and will be made, theories and frameworks that place anatomy in a framework alongside objective functions and learning rules offer a way to zero in on those features with the most explanatory power.

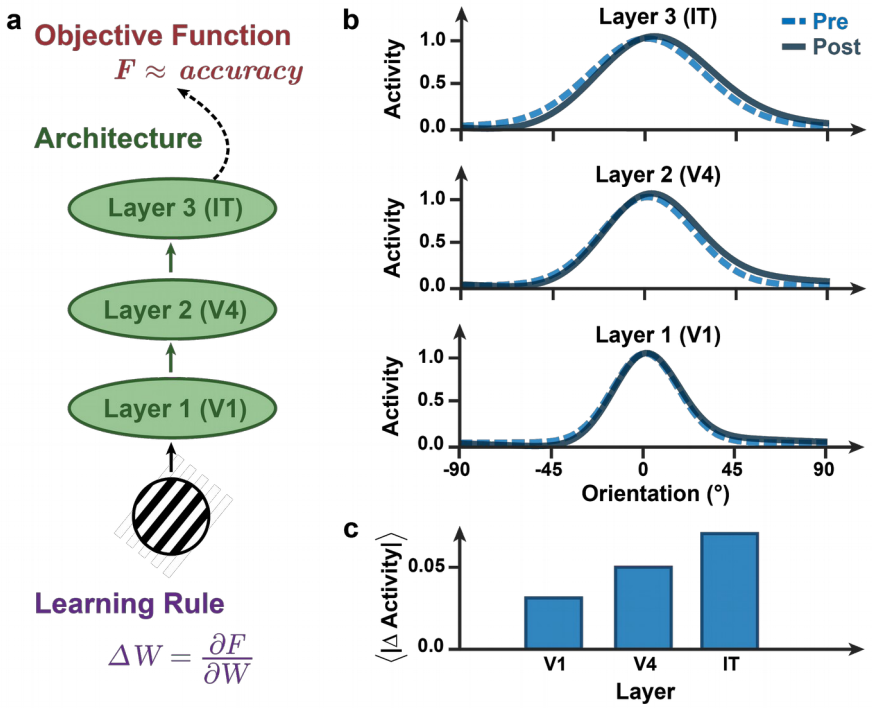

Figure 4: Comparing deep ANN models and the brain. One way to assess the three components at once is to compare experimental data with changes in representations in deep ANNs that incorporate all three components. **(a)** For example, we could use a deep ANN with a hierarchical architecture, trained with an objective function for maximizing rewards that are delivered when it successfully discriminates grating orientations, and a gradient-based, end-to-end learning rule. **(b)** When examining the orientation tuning of the populations in different layers of the hierarchy, such models can make predictions. For instance, the model may predict that the largest changes in tuning should occur higher in the cortical hierarchy (*top*), with smaller changes in the middle, e.g. in V4 (*middle*), and the smallest changes occurring low in the hierarchy, e.g. in V1 (*bottom*). **(c)** This leads to experimentally testable predictions about the average magnitude of changes in neural activity that should be observed experimentally when an animal is learning.

Empirical studies of learning rules in the brain

There is a long tradition in neuroscience of studying synaptic plasticity rules. Yet, these studies have rarely explored how credit assignment may occur. However, as we discussed above (Box 1), credit assignment is key to learning in ANNs, and may be in the brain as well. Thankfully, top-down feedback and neuromodulatory systems have become the focus of recent studies of synaptic plasticity[72–76]. This has allowed some concrete proposals, e.g. as to how apical dendrites may be involved in credit assignment[12,14], or how top-down attention mechanisms combined with neuromodulators may solve the credit assignment problem[37,38] (**Fig. 5**). We may also be able to look at changes in representations and infer the plasticity rules from those observations[77]. It is important for experimentalists to measure neural responses both during and after an animal has reached stable performance, so as to capture how representations evolve during learning. Work on learning rules with an eye to credit assignment is producing a finer-grained understanding of the myriad of factors that affect plasticity[78].

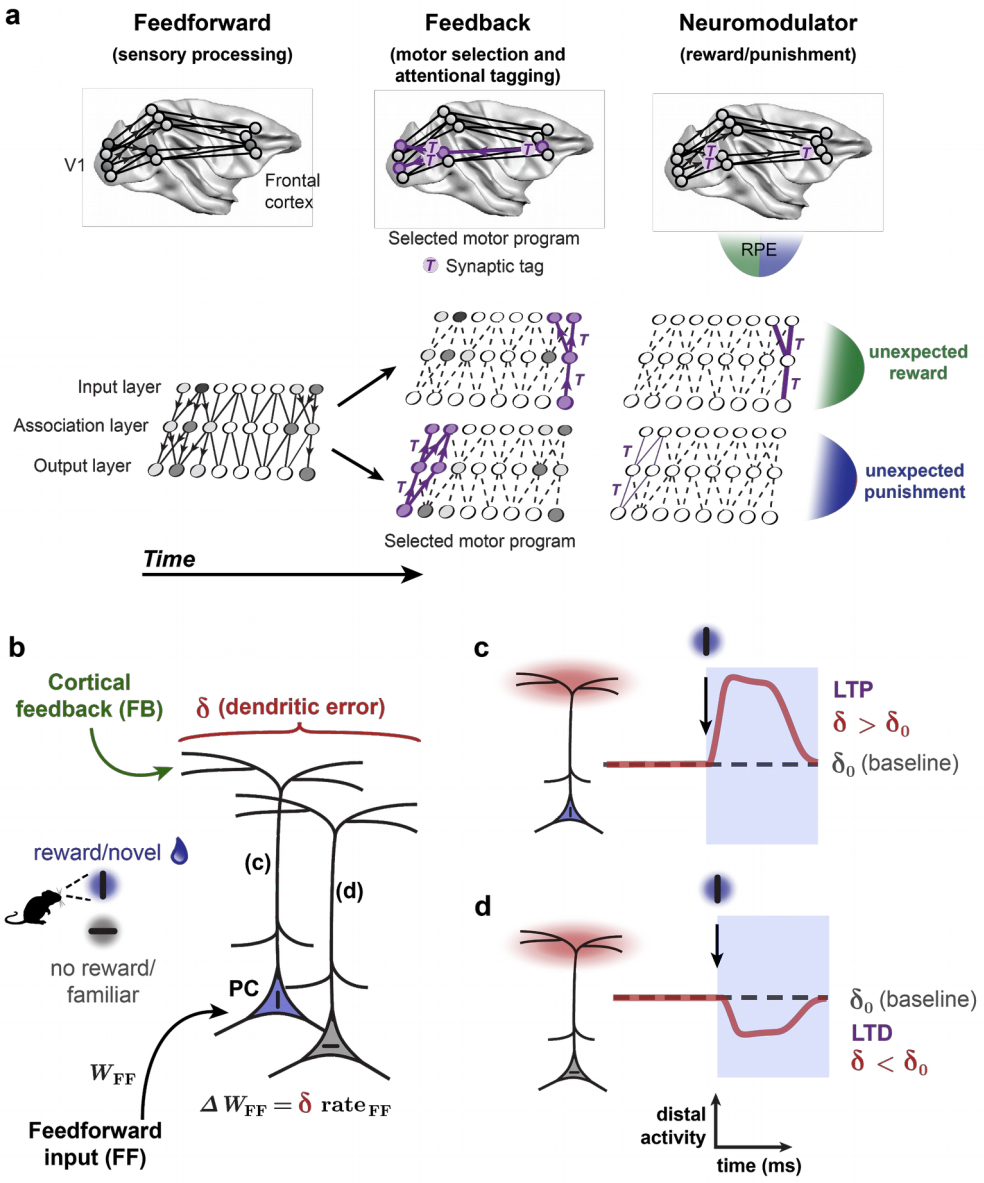

Figure 5: Biological models of credit assignment. (a) Attention based models of credit assignment (refs. 37,38) propose that the credit assignment problem is solved by the brain using attention and neuromodulatory signals. According to these models, sensory processing is largely feedforward in early stages, then feedback "tags" neurons and synapses for credit, and reward prediction errors (RPE) determine the direction of plastic changes. This is illustrated at the bottom, where circles indicate neurons, and the gray level indicates their level of activity. These models predict that the neurons responsible for activating a particular output unit will be tagged (T) by attentional feedback. Then, if a positive RPE is received, the synapses should potentiate. In contrast, if a negative RPE is received, the synapses should depress. This provides an estimate of a gradient for a category-based objective function. **(b-d)** Dendritic models of credit assignment (refs. 12,14) propose that gradient signals are carried by "dendritic error" (δ) signals in the apical dendrites of pyramidal neurons. **(b)** According to these models, feedforward weight updates are determined by a combination of feedforward inputs and δ. In an experiment where two different stimuli are presented, and only one is reinforced, this leads to specific predictions. **(c)** If a neuron is tuned towards a stimulus that is reinforced, then reinforcement should lead to an increase in apical activity. **(d)** In contrast, if a neuron is tuned to an unreinforced stimulus, its apical activity should decrease when reinforcement is received.

In the future, we should be better placed to study learning rules with optimization in mind. As optical technologies improve, and potentially give us a means of estimating synaptic changes *in vivo*[79], we may be able to directly relate synaptic changes to things like behavioral errors. We could also directly test hypothesized biological models of learning rules that can solve the credit assignment problem, such as those that use attention[37,38] or those that use dendritic signals for credit assignment[12,14] (**Fig. 5**).

Empirical studies of objective functions in the brain

In some cases, the objective functions being optimized by the brain may be represented directly in neural signals that we can monitor and record. In other cases, objective functions may only exist implicitly with respect to the plasticity rules that govern synaptic updates. Normative concepts, such as optimal control, are applicable[80], and evolutionary ideas can inform our thinking. More specifically, ethology may provide guidance[81] as to which functions would be useful for animals to optimize, giving us a meaningful intuitive space in which to think about objective functions.

There is a long-standing literature trying to relate experimental data to objective functions. This starts with theoretical work relating known plasticity rules to potential objective functions. For example, there are studies that attempt to estimate objective functions by comparing neural activity observed experimentally with the neural activity of ANNs trained on natural scenes[59,82]. There are also approaches that use inverse reinforcement learning to identify what a system optimizes[83]. Moreover, one could argue that we can get a handle on objective functions by looking for correlations between representational geometries optimized for a given objective and real neural representational geometries[28,84]. Another newly emerging approach asks what an animal's circuits can optimize when controlling a Brain Computer Interface (BCI) device[85]. Thus, a growing literature, which builds on previous work[80], helps us explore objective functions in the brain.

Caveats and concerns

One may argue that a focus on architectures, learning rules, and objective functions, and a move away from studying the coding properties of neurons, loses much of what we have learned so far, e.g. orientation selectivity, frequency tuning, spatial-tuning (place cells, grid cells). However, our proposed framework is heavily informed by this knowledge. Convolutional ANNs directly emerged from the observation of complex cells in the visual system[86]. Moreover, tuning curves are often measured in the context of learning experiments, and changes in tuning inform us about learning rules and objective functions.

In a similar vein, a lot of computational neuroscience has emphasized models of the dynamics of neural activity[87], and that has not been a major theme in our discussion. As such, one might worry that our framework fails to connect with this past literature. However, the framework we articulate here does not preclude consideration of dynamics. A focus on dynamics may equally be repurposed for making inferences about architectures, learning rules and objective functions, which have long been a feature of models of neural dynamics[49,88].

Another common objection to the relevance of deep learning for neuroscience is that many behaviors that animals engage in appear to require relatively little learning[48]. However, such innate behavior was "learned", only on evolutionary timescales. Hardwired behavior is,

arguably, best described as strong inductive biases, since even pre-wired behaviors can be modified by learning (e.g. horses still get better at running after birth). Hence, even when a neural circuit engages in only moderate amounts of learning, an optimization framework can help us model its operations[48].

The framework that we have laid out here makes the optimization of objective functions central to models of the brain. But a comprehensive theory of any brain likely requires attention to other constraints unrelated to any form of objective function optimization. For example, many aspects of physiology are determined by phylogenetic constraints that may be hold-overs from evolutionary ancestors. While these constraints are undoubtedly crucial for our models in neuroscience, we believe that it is the optimization of objective functions within these constraints that produces the rich diversity of neural circuitry and behavior that we observe in the brain.

Some of us, who are inclined to a bottom-up approach to understanding the brain, worry that attempts to posit objective functions or learning rules for the brain may be premature, needing far more details of brain operation than we currently possess. Nonetheless, scientific questions necessarily are posed within some framework of thought. Importantly, we are not calling for abandoning bottom-up explanations. Instead, we hope that important new experimental questions will emerge from the framework suggested by ANNs (see e.g. **Fig. 5**).

Finally, some researchers are concerned by the large number of parameters in deep ANNs, seeing them as a violation of Occam's razor and merely an overfitting to data. Interestingly, recent work in AI shows that the behavior of massively overparameterized learning systems can be counterintuitive—there appear to be intrinsic mathematical properties of over-parameterized learning systems that enable good generalization[42,89]. Since the brain itself apparently contains a massive number of potential parameters to adapt (e.g. synaptic connections, dendritic ion channel densities, etc.), one might argue that the large number of parameters in deep ANNs actually makes them even more appropriate models of the brain.

Conclusion

Much of systems neuroscience has attempted to formulate succinct statements about the function of individual neurons in the brain. This approach has been successful at explaining some (relatively small) circuits and certain hard-wired behaviors. However, there is reason to believe that this approach will need to be complemented by other insights if we are to develop good models of plastic circuits with thousands, millions or billions of neurons. There is, unfortunately, no guarantee that the function of individual neurons in the central nervous system can be compressed down to a human-interpretable, verbally articulable form. Given that we currently have no good means of distilling the function of individual units in deep ANNs into words, and given that real brains are likely more, not less, complex, we suggest that systems neuroscience would benefit from focusing on the kinds of models that have been successful in ANN research programs, i.e. models grounded in the three essential components.

Current theories in systems neuroscience are beautiful and insightful, but we believe that they could benefit from a cohesive framework founded in optimization. For example, local plasticity rules, such as Hebbian mechanisms, explain a great deal of biological data. But, to achieve good performance on complex tasks, Hebbian rules must be designed with objective functions and architectures in mind[34,90]. Similarly, other researchers have, for good reason, pointed out the benefits of the inductive biases utilized by the brain[48]. However, inductive biases

are not on their own sufficient to solve complex tasks, like those contained in the AI Set or various Brain Sets. To solve these difficult problems, inductive biases must be paired with learning and credit assignment. If, as we have argued, the set of tasks that an animal can solve are an essential consideration for neuroscience, then it is critical to build models that can actually solve these tasks.

Inevitably, both bottom-up descriptive work and top-down theoretical work will be required to make progress in systems neuroscience. It is important, though, to start with the right kind of top-down theoretical framing. Given the ability of modern machine learning to solve problems in the AI Set and numerous Brain Sets, it will be fruitful to guide the top-down framework of systems neuroscience research with machine learning insights. If we consider research data within the framework provided by this mindset, and focus our attention on the three essential components identified here, we believe we can develop theories of the brain that will reap the full benefits of the current technological revolution in neuroscience.

Acknowledgements
This article emerged from a workshop on optimization in the brain that happened February 24-28, 2019 at the Bellairs Research Institute of McGill University. We would like to thank Element AI and Bellairs Research Institute for their critical support in organizing this workshop.

References

1. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
2. Steinmetz, N. A., Koch, C., Harris, K. D. & Carandini, M. Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. *Neurotechnologies* **50**, 92–100 (2018).
3. Marder, E. & Bucher, D. Central pattern generators and the control of rhythmic movements. *Curr. Biol.* **11**, R986–R996 (2001).
4. Cullen, K. E. The vestibular system: multimodal integration and encoding of self-motion for motor control. *Trends Neurosci.* **35**, 185–196 (2012).
5. Kim, J. S. *et al.* Space–time wiring specificity supports direction selectivity in the retina. *Nature* **509**, 331 (2014).
6. Olshausen, B. A. & Field, D. J. What is the other 85 percent of V1 doing. *Van Hemmen T Sejnowski Eds* **23**, 182–211 (2006).
7. Thompson, L. & Best, P. Place cells and silent cells in the hippocampus of freely-behaving rats. *J. Neurosci.* **9**, 2382–2390 (1989).
8. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19**, 356–365 (2016).
9. Botvinick, M. *et al.* Reinforcement Learning, Fast and Slow. *Trends Cogn. Sci.* (2019).
10. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. Neurosci.* 1 (2018).
11. Rumelhart, D. E., McClelland, J. L. & PDP Research Group. *Parallel distributed processing*. **1**, (MIT press Cambridge, 1988).
12. Sacramento, J., Costa, R. P., Bengio, Y. & Senn, W. Dendritic cortical microcircuits approximate the backpropagation algorithm. in 8735–8746 (2018).
13. Poirazi, P., Brannon, T. & Mel, B. W. Pyramidal Neuron as Two-Layer Neural Network. *Neuron* **37**, 989–999 (2003).
14. Guerguiev, J., Lillicrap, T. P. & Richards, B. A. Towards deep learning with segregated dendrites. *eLife* **6**, e22901 (2017).
15. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (MIT press, 2016).
16. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).
17. Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98**, 630-644.e16 (2018).
18. Richards, B. A. & Lillicrap, T. P. Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* **54**, 28–36 (2019).
19. Roelfsema, P. R. & Holtmaat, A. Control of synaptic plasticity in deep cortical networks. *Nat. Rev. Neurosci.* **19**, 166 (2018).
20. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep

convolutional neural networks. in 1097–1105 (2012).

21. Hannun, A. *et al.* Deep speech: Scaling up end-to-end speech recognition. *ArXiv Prepr. ArXiv14125567* (2014).
22. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 8 (2019).
23. Gao, Y., Hendricks, L. A., Kuchenbecker, K. J. & Darrell, T. Deep learning for tactile understanding from visual and haptic data. in 536–543 (IEEE, 2016).
24. Banino, A. *et al.* Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018).
25. Finn, C., Goodfellow, I. & Levine, S. Unsupervised learning for physical interaction through video prediction. in 64–72 (2016).
26. Silver, D. *et al.* Mastering the game of go without human knowledge. *Nature* **550**, 354 (2017).
27. Santoro, A. *et al.* A simple neural network module for relational reasoning. in 4967–4976 (2017).
28. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput Biol* **10**, e1003915 (2014).
29. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
30. Pospisil, D. A., Pasupathy, A. & Bair, W. 'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. *eLife* **7**, e38242 (2018).
31. Singer, Y. *et al.* Sensory cortex is optimized for prediction of future input. *eLife* **7**, e31557 (2018).
32. Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M. & Tanaka, K. Illusory Motion Reproduced by Deep Neural Networks Trained for Prediction. *Front. Psychol.* **9**, 345 (2018).
33. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
34. Scellier, B. & Bengio, Y. Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Front. Comput. Neurosci.* **11**, 24 (2017).
35. Whittington, J. C. & Bogacz, R. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Comput.* (2017).
36. Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* **7**, 13276 (2016).
37. Roelfsema, P. R. & Ooyen, A. van. Attention-Gated Reinforcement Learning of Internal Representations for Classification. *Neural Comput.* **17**, 2176–2214 (2005).
38. Pozzi, I., Bohté, S. & Roelfsema, P. A Biologically Plausible Learning Rule for Deep Learning in the Brain. *ArXiv Prepr. ArXiv181101768* (2018).
39. Körding, K. P. & König, P. Supervised and Unsupervised Learning with Two Sites of Synaptic Integration. *J. Comput. Neurosci.* **11**, 207–215 (2001).
40. Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an Integration of Deep Learning and Neuroscience. *Front. Comput. Neurosci.* **10**, (2016).

41. Raman, D. V., Rotondo, A. P. & O'Leary, T. Fundamental bounds on learning performance in neural circuits. *Proc. Natl. Acad. Sci.* 201813416 (2019).

42. Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y. & Srebro, N. The role of over-parametrization in generalization of neural networks. (2018).

43. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).

44. Bengio, Y. & LeCun, Y. Scaling learning algorithms towards AI. *Large-Scale Kernel Mach.* **34**, 1–41 (2007).

45. Neyshabur, B., Tomioka, R. & Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *ArXiv Prepr. ArXiv14126614* (2014).

46. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. in 4077–4087 (2017).

47. Ravi, S. & Larochelle, H. Optimization as a model for few-shot learning. (2016).

48. Zador, A. M. A Critique of Pure Learning: What Artificial Neural Networks can Learn from Animal Brains. *Nat. Commun.* **10**, 1–7 (2019).

49. Bellec, G., Salaj, D., Subramoney, A., Legenstein, R. & Maass, W. Long short-term memory and learning-to-learn in networks of spiking neurons. in 787–797 (2018).

50. Huang, Y. & Rao, R. P. N. Predictive coding. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 580–593 (2011).

51. Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**, 229–256 (1992).

52. Klyubin, A. S., Polani, D. & Nehaniv, C. L. Empowerment: A universal agent-centric measure of control. in **1**, 128–135 (IEEE, 2005).

53. Salge, C., Glackin, C. & Polani, D. Empowerment–an introduction. in *Guided Self-Organization: Inception* 67–114 (Springer, 2014).

54. Newell, A. & Simon, H. A. *GPS, a program that simulates human thought*. (RAND CORP SANTA MONICA CALIF, 1961).

55. Nguyen, A., Yosinski, J. & Clune, J. Understanding Neural Networks via Feature Visualization: A survey. *ArXiv Prepr. ArXiv190408939* (2019).

56. Kebschull, J. M. *et al.* High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA. *Neuron* **91**, 975–987 (2016).

57. Kornfeld, J. & Denk, W. Progress and remaining challenges in high-throughput volume electron microscopy. *Curr. Opin. Neurobiol.* **50**, 261–267 (2018).

58. Lillicrap, T. P. & Kording, K. P. What does it mean to understand a neural network? *ArXiv Prepr. ArXiv190706374* (2019).

59. Olshausen, B. A. & Field, D. J. Natural image statistics and efficient coding. *Netw. Comput. Neural Syst.* **7**, 333–339 (1996).

60. Hyvärinen, A. & Oja, E. Simple neuron models for independent component analysis. *Int. J. Neural Syst.* **7**, 671–687 (1996).

61. Oja, E. Simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**, 267–273 (1982).

62. Intrator, N. & Cooper, L. N. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Netw.* **5**, 3–17 (1992).

63. Fiser, A. *et al.* Experience-dependent spatial expectations in mouse visual cortex. *Nat Neurosci* **advance online publication**, (2016).

64. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997).

65. Momennejad, I. *et al.* The successor representation in human reinforcement learning. *bioRxiv* (2016). doi:10.1101/083824

66. Nayebi, A. *et al.* Task-Driven convolutional recurrent models of the visual system. in 5290–5301 (2018).

67. Schrimpf, M. *et al.* Brain-Score: which artificial neural network for object recognition is most brain-like? *BioRxiv* 407007 (2018).

68. Kepecs, A. & Fishell, G. Interneuron cell types are fit to function. *Nature* **505**, 318–326 (2014).

69. Van Essen, D. C. & Anderson, C. H. Information processing strategies and pathways in the primate visual system. *Introd. Neural Electron. Netw.* **2**, 45–76 (1995).

70. Lindsey, J., Ocko, S. A., Ganguli, S. & Deny, S. A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. *ArXiv Prepr. ArXiv190100945* (2019).

71. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).

72. Kwag, J. & Paulsen, O. The timing of external input controls the sign of plasticity at local synapses. *Nat. Neurosci.* **12**, 1219 (2009).

73. Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S. & Magee, J. C. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033 (2017).

74. Lacefield, C. O., Pnevmatikakis, E. A., Paninski, L. & Bruno, R. M. Reinforcement Learning Recruits Somata and Apical Dendrites across Layers of Primary Sensory Cortex. *Cell Rep.* **26**, 2000–2008 (2019).

75. Williams, L. E. & Holtmaat, A. Higher-Order Thalamocortical Inputs Gate Synaptic Long-Term Potentiation via Disinhibition. *Neuron* (2019). doi:10.1016/j.neuron.2018.10.049

76. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616 (2014).

77. Lim, S. *et al.* Inferring learning rules from distributions of firing rates in cortical neurons. *Nat Neurosci* **18**, 1804–1810 (2015).

78. Costa, R. P. *et al.* Synaptic transmission optimization predicts expression loci of long-term plasticity. *Neuron* **96**, 177–189 (2017).

79. Zolnik, T. A. *et al.* All□optical functional synaptic connectivity mapping in acute brain slices using the calcium integrator CaMPARI. *J. Physiol.* **595**, 1465–1477 (2017).

80. Scott, S. H. Optimal feedback control and the neural basis of volitional motor control. *Nat. Rev. Neurosci.* **5**, 532 (2004).

81. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).

82. Zylberberg, J., Murphy, J. T. & DeWeese, M. R. A Sparse Coding Model with Synaptically Local Plasticity and Spiking Neurons Can Account for the Diverse Shapes of V1 Simple Cell Receptive Fields. *PLOS Comput. Biol.* **7**, e1002250 (2011).

83. Chalk, M., Tkačik, G. & Marre, O. Inferring the function performed by a recurrent neural network. *bioRxiv* 598086 (2019). doi:10.1101/598086

84. Cadieu, C. F. *et al.* Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Comput Biol* **10**, e1003963 (2014).

85. Golub, M. D. *et al.* Learning by neural reassociation. *Nat. Neurosci.* **21**, 607–616 (2018).

86. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).

87. Vogels, T. P., Rajan, K. & Abbott, L. F. Neural Network Dynamics. *Annu. Rev. Neurosci.* **28**, 357–376 (2005).

88. Koren, V. & Denève, S. Computational Account of Spontaneous Activity as a Signature of Predictive Coding. *PLOS Comput. Biol.* **13**, e1005355 (2017).

89. Advani, M. S. & Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *ArXiv Prepr. ArXiv171003667* (2017).

90. Amit, Y. Deep learning with asymmetric connections and Hebbian updates. *Front. Comput. Neurosci.* **13**, (2019).

91. Samadi, A., Lillicrap, T. P. & Tweed, D. B. Deep learning with dynamic spiking neurons and fixed feedback weights. *Neural Comput.* **29**, 578–602 (2017).

92. Akrout, M., Wilson, C., Humphreys, P. C., Lillicrap, T. & Tweed, D. Using Weight Mirrors to Improve Feedback Alignment. *ArXiv Prepr. ArXiv190405391* (2019).

93. Lansdell, B. & Kording, K. Spiking allows neurons to estimate their causal effect. *bioRxiv* 253351 (2018).

94. Werfel, J., Xie, X. & Seung, H. S. Learning curves for stochastic gradient descent in linear feedforward networks. in 1197–1204 (2004).

95. Bartunov, S. *et al.* Assessing the scalability of biologically-motivated deep learning algorithms and architectures. in 9368–9378 (2018).

96. MacKay, D. J. *Information theory, inference and learning algorithms*. (Cambridge university press, 2003).

97. Goel, V., Weng, J. & Poupart, P. Unsupervised video object segmentation for deep reinforcement learning. in 5683–5694 (2018).

98. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **3361**, 1995 (1995).

99. Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. & Bengio, Y. Attention-based models for speech recognition. in 577–585 (2015).

100. Houthooft, R. *et al.* Vime: Variational information maximizing exploration. in 1109–1117 (2016).